SHORT COMMUNICATION

Mapping microsatellite markers identified in porcine EST sequences[1]

G. A. Rohrer[*], S. C. Fahrenkrug[*], D. Nonneman[*], N. Tao[†] and W. C. Warren[†]

[*]USDA, ARS, US Meat Animal Research Center, Clay Center, NE, USA. [‡]Monsanto, 700 Chesterfield Parkway North, Chesterfield, MO, USA

Summary A sequence search of swine expressed sequence tags (EST) data in GenBank identified over 100 sequence files which contained a microsatellite repeat or simple sequence repeat (SSR). Most of these repeat motifs were dinucleotide (CA/GT) repeats; however, a number of tri-, tetra-, penta- and hexa-nucleotide repeats were also detected. An initial assessment of six dinucleotide and 14 higher-order repeat markers indicated that only dinucleotide markers yielded a sufficient number of informative markers (100% vs. 14% for dinucleotide and higher order repeats, respectively). Primers were designed for an additional 50 di- and one tri-nucleotide SSRs. Overall, 42 markers were polymorphic in the US Meat Animal Research Center (MARC) reference population, 17 markers were uninformative and 12 primer pairs failed to satisfactorily amplify genomic DNA. A comparison of di-nucleotide repeat vs. markers with repeat motifs of three to six bases demonstrated that 72% of dinucleotide markers were informative relative to only 7% of other repeat motifs. The difference was the result of a much higher percentage of monomorphic markers in the three to six base repeat motif markers than in the dinucleotide markers (64% vs. 14%). Either higher order repeat motifs are less polymorphic in the porcine genome or our selection criteria for repeat length of more than 17 contiguous bases was too low. The mapped microsatellite markers add to the porcine genetic map and provide valuable links between the porcine and human genome.

Keywords chromosome, EST, map, microsatellite, porcine, SSR.

A large number of single-pass sequence data [expressed sequence tags (EST)] of cDNA clones from numerous swine tissues have been collected by many laboratories (Seo & Beever 2001; Tuggle *et al.* 2001; Fahrenkrug *et al.* 2002). These sequence data are often between 400 and 600 bases in length. More than 100 000 porcine EST sequences now exist in public databases as summarized in The Institute for Genomic Research's (TIGR) database (Quackenbush *et al.* 2001; http://www.tigr.org/tdb/ssgi/, v4.0) and they are assembled into 47 540 unique sequences. A BLAST search of tentative contigs (TC) and other EST sequence data in the TIGR database indicates that a large number of sequences contain microsatellite or simple sequence repeat (SSR), motifs. Because one of the current objectives of the US Meat Animal Research Center (MARC) Swine Mapping group is to place a significant number of ESTs on the porcine genetic map, the objective of the current study was to determine the utility of microsatellite markers developed from EST sequences when mapping known genes.

Our initial attempt to identify SSR in EST sequences was prior to the first release of the TIGR porcine gene index. The GenBank database was searched to identify porcine EST sequences which contained microsatellite repeats with

Address for correspondence

Gary A. Rohrer, USDA, ARS, US Meat Animal Research Center, Spur 18D, PO Box 166, Clay Center, NE 68933-0166, USA.
E-mail: rohrer@email.marc.usda.gov

Present address: S. C. Fahrenkrug, Present address: University of Minnesota, 1988 Fitch Ave., St Paul, MN 55 108, USA.

Present address: W. C. Warren, Washington University School of Medicine, 4444 Forest Park Blvd., St Louis, MO 63 108, USA.

Accepted for publication 11 March 2002

[1] This article is the material of the US Government, and can be produced by the public at will.

motifs ranging from two to six bases using software developed by Tao and coworkers (unpublished). To standardize the cutoff for the number of repeats necessary to evaluate the marker, it was decided that all SSR that spanned more than 17 contiguous bases would be tested. From this search 20 amplicons were developed for the assessment of repeat motifs. These amplicons contained di-($n = 6$), tri-($n = 7$), tetra-($n = 1$), penta-($n = 1$) and hexa-($n = 4$) nucleotide repeat motifs and one amplicon (*ESTMS17*) contained a compound repeat motif of $(GCT)_5$ $C(GA)_9$ $(CAGA)_2$ (Table 1). These markers were designated as *ESTMS1-21* (it was determined that one dinucleotide marker, *ESTMS18*, had already been mapped and was removed from this study). Primers were selected based on PRIMER 3.0 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi), with targeted amplicon size less than 150 bases in order to minimize the probability of primers spanning intron/exon junctions.

Evaluation of the initial 20 amplicons determined that only dinucleotide repeat motifs yielded a sufficient number of informative markers. Once the TIGR porcine gene index was available, the next search utilized the BLAST procedure at the TIGR website, primarily targeting CA/GT microsatellites. The search resulted in 246 sequences which met the criteria for repeat length (0.5% of unique sequences). From these sequences, 51 amplicons were developed and contained di-($n = 50$) or tri-($n = 1$) nucleotide repeat motifs (Table 1). Selection criteria for genetic marker development was based on their assignment to a tentative orthologue group (TOGA) or quality of the sequence for marker development (number of repeat units and amount of sequence available 5′ and 3′ of the repeat). For TCs which were assigned a gene name, the markers were named after the acronym of the gene. The remaining markers were assigned an SE prefix followed by their MARC sequence ID number (same as TIGR's EST ID) because all of these markers were developed from the MARC EST libraries (Fahrenkrug *et al.* 2002).

Amplicons were optimized and genotyped across seven families (86 progeny) of the MARC swine reference population as described by Rohrer *et al.* (1996). Linkage analyses were conducted as described (Rohrer *et al.* 1996) where TWOPOINT analyses were used to indicate the chromosome linkage group and the ALL, FLIPS and FIXED options were used to determine marker position (CRIMAP v2.4; Green *et al.* 1990). Suspect genotypes, as determined by CHROMPIC, were checked and all appropriate corrections made. Multipoint locations for all mapped markers were based on the last published MARC swine genetic map (Rohrer *et al.* 1996; http://www.marc.usda.gov/).

Table 1 presents success rates for the different types of microsatellite repeat motifs. Overall, only 59% of amplicons developed were informative in the MARC reference family. A large difference existed between percentage of informative markers developed from dinucleotide repeats vs. repeat motifs ranging from three to six bases in length (72% vs. 7%, respectively). Percentage of amplicons which failed to amplify was twice as great, and percentage of uninformative amplicons was more than four times higher for amplicons derived from repeat motifs of three to six bases than dinucleotide repeats (Table 1). Because these failure rates are based on a very small number of amplicons, some of these differences could be due to chance. However, the higher percentage of monomorphic PCR products from amplicons containing repeat motifs of three to six bases is probably real. Some of the tri- and/or hexa-nucleotide repeats may have been within coding regions, and therefore, repeat number within the coding region was not polymorphic. Furthermore, amplicons within coding regions would be more likely to contain an intron, and the amplicon would then have been coded as having no amplified product of the appropriate size.

Among the seven trinucleotide repeats, three motifs were observed twice (CTG, CAA, and GGC) and AGG was observed once. The number of repeat units ranged from 6 to 9. *ESTMS4* repeat motif was $(AGG)_8$. All four hexanucleotide repeats were unique and contained three repeats. However, GGC was present in three of the four hexanucleotide motifs.

Comparison of results from dinucleotide markers in this study with results obtained at MARC in a larger effort to isolate anonymous genomic CA/GT repeats (Alexander *et al.*

Table 1 Performance of amplicons which amplify microsatellite regions identified in EST sequences.

Repeat type	Number of amplicons	Failed to amplify	Monomorphic product	Polymorphic product
2	56	8 (14)[1]	8 (14)	40 (72)
3	8	3 (38)	4 (50)	1 (12)
4–6	6	1 (17)	5 (83)	0 (0)
Compound	1	0 (0)	0 (0)	1 (100)
Total	71	12 (17)	17 (24)	42 (59)
Anonymous dinucleotides[2]	1036	115 (11)	39 (4)	882 (85)

[1]Numbers in parentheses are percentages of the number of amplicons.

[2]Dinucleotide repeats isolated from anonymous genomic fragments (Alexander *et al*. 1996).

1996) reveals that a similar percentage of these markers were successfully amplified ($P > 0.45$; based on a chi-square test); however, significantly more amplicons in the present study were uninformative (14% vs. 4%; $P < 0.0001$). Although the threshold for minimum number of repeats was the same for both studies (more than 17 contiguous bases in this study and in Alexander *et al.* 1996), the average contiguous repeats present in EST derived sequences was less than in anonymous genomic fragments (12.9 vs. 17.3 uninterrupted repeats). Similarly, fewer alleles were observed in the EST derived markers than in the anonymous markers (4.3 vs. 5.6 alleles, respectively). Because most of the MARC EST sequence templates were PCR products (Smith *et al.* 2000; Fahrenkrug *et al.* 2002), it is possible that the longer microsatellites would not have yielded enough high quality sequence data after the repeat regions to design primers or be assembled into TCs. Another plausible explanation might be that CA/GT microsatellites have fewer repeats when residing in UTRs than when in intervening DNA. Figure 1 depicts a plot of number of repeats vs. number of alleles for EST and anonymous dinucleotide repeats. Both data sets indicate that as the number of repeats increase so does the number of alleles. In the present study, the average number of repeats was less in the uninformative dinucleotide markers than the informative dinucleotide markers (11.6 vs. 13.3, respectively).

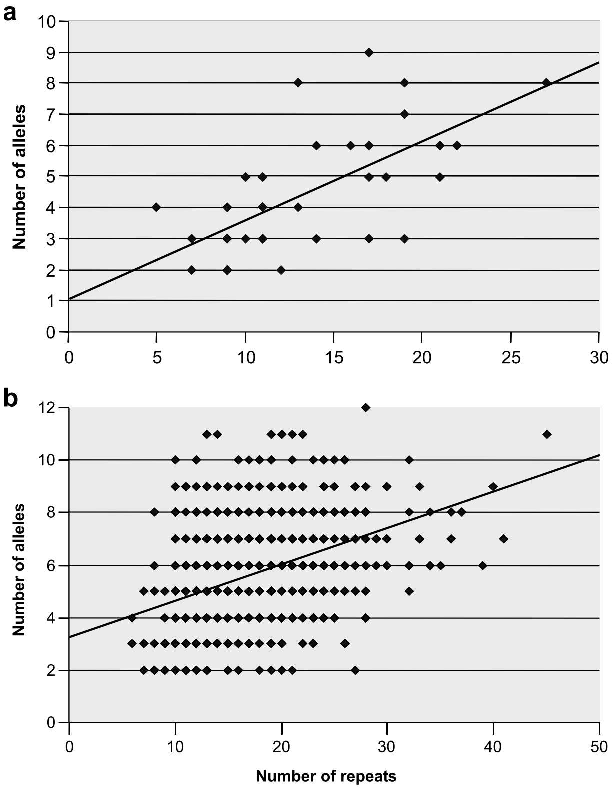

Figure 1 Scatter plot of number of alleles vs. number of repeats for informative dinucleotide repeat motif markers. Solid line indicates the least-squares regression line. (a) Markers derived from EST sequences. Best fit line had an intercept of 1.0 and a slope of 0.25. (b) Markers derived from an anonymous approach (Alexander *et al.* 1996). Best fit line had an intercept of 3.3 and a slope of 0.14.

Information for all markers which were informative in the MARC swine reference population is presented in Table 2. While 42 markers were informative in the MARC Swine Reference Population, only 41 markers were placed in the genetic map as one marker had too few meioses to detect significant two-point linkages (*ESTMS4*). Ironically, this marker was the only informative trinucleotide repeat motif investigated in this study. Markers were assigned to 16 of 19 porcine chromosomes. Six chromosomes had one EST marker (SSC1, 2, 11, 13, 14 and 17), two had two EST markers (SSC10 and 15), four had three markers (SSC7, 8, 12, and X), two had four EST markers (SSC6 and 9), while SSC4 had five EST markers and SSC3 had six EST markers. Twenty-eight (67%) mapped EST sequences were associated with TCs in the TIGR gene index. Twenty-five TC associated markers, as well as three additional markers, had a predicted orthologous position in the human genome (December 2001 build at http://genome.ucsc.edu/). It was expected *a priori* that few of these markers would have identified orthologues because the EST sequences were short and most of the microsatellite sequences reside in 3′ UTR which are not well conserved across species. However, as more EST sequences are collected and more complete cDNA sequence becomes available, most of these markers should be assigned to known transcripts and/or orthologous positions in the human genome. To demonstrate this point only 44% of the current markers had orthologous human positions in TIGR's version 3.0 of the porcine gene index and 24 markers were in TCs (approximately 85 000 EST sequences). When version 4.0 (based on 100 000 EST sequences) is used 67% of the markers have orthologous human positions and 28 markers are in TCs.

All but three of the markers with a mapped human orthologue were mapped to expected locations based on bi-direction chromosomal painting (Goureau *et al.* 1996). One exception was marker *SE47351*, which mapped to SSC3 position 14. This marker is orthologous to a leucine rich neuronal protein mapped to HSA7q22. Goureau *et al.* (1996) did not identify homology between HSA7q and SSC3p. However, erythropoietin (Liu *et al.* 1998), β-actin (Thomsen *et al.* 1998) and cytochrome P450 subfamily IIIA (Thomsen *et al.* 1998) all reside on HSA7q and have been mapped to SSC3p. Our information supports a conserved syntenic unit present on SSC3 and HSA7q. *SE47346*, which mapped to SSC15, presented another dilemma. While TIGR had assigned the contig to a gene mapped to HSAX, a

Table 2 Information on the EST clones which possessed informative microsatellite markers.

Marker[1]	TC[2]	Number of alleles[3]	Genomic position[4]	Gene acronym[5]	Human position[6]
AIP4	40 297	9	12 : 2	*BIRC5*	17 : 79.9
ESTMS1		6	6 : 77.5		
ESTMS16		4	9 : 4.8		
ESTMS17	43 918	4	3 : 92		2 : 25.7
ESTMS19		2	3 : 44		
ESTMS2		6	9 : 85		
ESTMS20		8	9 : 53		
ESTMS21	41 095	3	6 : 93	*UP*	1 : 37.8
ESTMS4		2			
GLS	34 302	2	15 : 67	*GLS*	2 : 190.2
PC3	40 206	4	3 : 59	*ACTR1B*	2 : 94.4
PIP5K2	41 036	4	10 : 91	*PIP5K2 A*	10 : 23.3
RAB7	33 221	5	13 : 57	*RAB7*	3 : 133.5
RIGB	32 585	2	2 : 28	*AP1G1*	11 : 66.4
SE13123	42 374	4*	X : 74		20 : 36.0
SE15078	41 931	6*	X : 74		X : 49.3
SE234324	40 400	2	7 : 99	*NUMB*	14 : 72.3
SE259162	36 672	8	12 : 103	*SPAG7*	17 : 5.2
SE27951	35 177	5	1 : 57		6 : 102.9
SE28540	33 722	3	6 : 97	*KIAA1135*	18 : 13.3
SE29505	32 010	3	11 : 90		
SE30433		5	10 : 43		
SE45980		3	4 : 66		
SE47323		3	7 : 86		15 : 30.0
SE47329	32 240	8	3 : 43	*KIAA1080*	16 : 24.3
SE47346	32 566	6	15 : 56		8 : 7.1; X : 130.8[7]
SE47349		5	6 : 89.3	*WNT4*	1 : 22.4
SE47350		6	4 : 131		
SE47351	43 069	4	3 : 14		7 : 102.8
SE47381	39 741*	4	4 : 121	*NES*	1 : 156.3
SE47407	45 952	2	9 : 113		1 : 181.0
SE47408		3	4 : 79		1 : 157.1
SE47610	44 211	7	8 : 68		
SE47615	39 786	3	7 : 70	*SLC29A1*	6 : 47.5
SE47623	41 342	4	17 : 89	*UBE2L6*	20 : 56.9
SE47648	37 829	6	8 : 57		
SE47656	41 114	3	4 : 77	*KIAA1355*	1 : 160.7
SE47665	45 939	3*	X : 74	*ZNF261*	X : 65.7
SE51018		3	3 : 19		
SE77660		5	14 : 72		
SE77921	39 244	2	12 : 52	*GRN*	17 : 44.8
WFS	44 139	3	8 : 9	*WFS1*	4 : 7.3

[1] Details of markers including primer sequences can be found at http://www.marc.usda.gov.

[2] TC is the tentative contig assigned in TIGR's porcine gene index version 4.0. An asterisk (*) indicates that the microsatellite was identified from sequences derived from the 3′ end of the cDNA clones, but the TC was assigned based on sequence from the 5′ end of the same cDNA clone.

[3] Number of alleles marked with an asterisk (*) indicates that one of the detected alleles was a null allele.

[4] Genomic position is presented as the porcine chromosome : cM position in the MARC swine genetic map (http://www.marc.usda.gov/).

[5] Gene acronym is the primary acronym for the human orthologue as detected in the Genomic Database (GDB; http://www.gdb.org/).

[6] Human position is presented as chromosome : megabase position as presented at (http://genome.ucsc.edu/) December 2001 build.

[7] TIGR gene index indicates the location of the human orthologue resides at HSAX : 130.8. However, the most significant sequence identified in a BLAT search was located at HSA8 : 7.1.

BLAST search of the contig sequence against the human genomic sequence revealed the best match to a location on HSA8 approximate position 7.1 Mb (December 2001 build, http://genome.ucsc.edu/). While Goureau *et al.* (1996) did not identify an orthologous segment between HSA8 and SSC15, Wintero *et al.* (1998) have mapped a gene (*HSEF1B*) located on HSA8 to SSC15. Because it is unlikely that the true orthologue to *SE47346* actually maps to HSAX, it is presumed that the position on HSA8 actually is the orthologous region to SSC15. Finally, the only homology to *SE13123* detected in the human genome was on HSA20. As *SE13123* maps to SSCX, it is doubtful that this is the correct ortholog in the human genome.

Developing microsatellite markers from dinucleotide repeats residing in EST sequences appears to be an effective approach to placing expressed genes in the porcine genetic map. More than 70% of amplicons developed for dinucleotides were mapped and the cost to map these ESTs was equal to or less than any other methodology currently available. Few higher order repeat motifs were found in porcine EST sequences, and the performance of markers developed from higher order repeats in this study was unsatisfactory. Therefore, development of genetic markers for ESTs containing SSRs with repeat motifs greater than two bases should focus on mapping techniques which do not require variations in length.

Acknowledgements

The authors wish to acknowledge K. Simmerman and S. Freston for technical assistance and S. Kluver for manuscript preparation.

References

Alexander, L.J., Rohrer, G.A. & Beattie, C.W. (1996) Cloning and characterization of 414 polymorphic porcine microsatellites. *Animal Genetics* **27**, 137–48.

Fahrenkrug, S.C., Smith, T.P.L., Freking, B.A. *et al.* (2002) Porcine gene discovery by normalized cDNA-library sequencing and EST cluster assembly. *Mammalian Genome* (in press).

Goureau, A., Yerle, M., Schmitz, A., Riquet, J., Milan, D., Pinton, P., Frelat, G. & Gellin, J. (1996) Human and porcine correspondence of chromosome segments using bidirectional chromosome painting. *Genomics* **36**, 252–62.

Green, P., Falls, K. & Crooks, S. (1990) *Documentation for CRI-MAP*, Version 2.4. Washington University School of Medicine. St Louis, MO.

Liu, W.S., Harbitz, I., Gustavsson, I. & Chowdhary, B.P. (1998) Mapping of the porcine erythropoietin gene to chromosome 3p15-p16 and ordering of four related subclones by Fiber-FISH and DNA-combing. *Hereditas* **128**, 77–81.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. & White, J. (2001) The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research* **29**, 159–64.

Rohrer, G.A., Alexander, L.J., Hu, Z., Smith, T.P.L., Keele, J.W. & Beattie, C.W. (1996) A comprehensive map of the porcine genome. *Genome Research* **6**, 371–91.

Seo, K. & Beever, J.E. (2001) Monitoring gene expression in swine skeletal muscle. *Proceedings of the Plant and Animal Genome IX, San Diego, CA, Abstract.* **W207**, 59.

Smith, T.P.L., Godtel, R.A. & Lee, R.T. (2000) PCR-based reaction setup for high-throughput cDNA library sequencing on the ABI 3700 automated DNA sequencer. *Biotechniques* **29**, 698–700.

Thomsen, P.D., Winterø, A.K. & Fredholm, M. (1998) Chromosomal assignments of 19 porcine cDNA sequences by FISH. *Mammalian Genome* **9**, 394–6.

Tuggle, C.K., Green, J., Fitzsimmons, C.J., Woods, R., Prather, P. *et al.* (2001) Development of resources for functional genomics in the pig, production of 14 cDNA libraries and sequencing of over 7,000 clones from female reproductive tissues. *Proceedings of the Plant and Animal Genome IX, San Diego, CA, Abstract.* **P67**, 77.

Wintero, A.K., Jorgensen, C.B., Robic, A., Yerle, M. & Fredholm, M. (1998) Improvement of the porcine transcription map: localization of 33 genes, of which 24 are orthologous. *Mammalian Genome* **9**, 366–72.